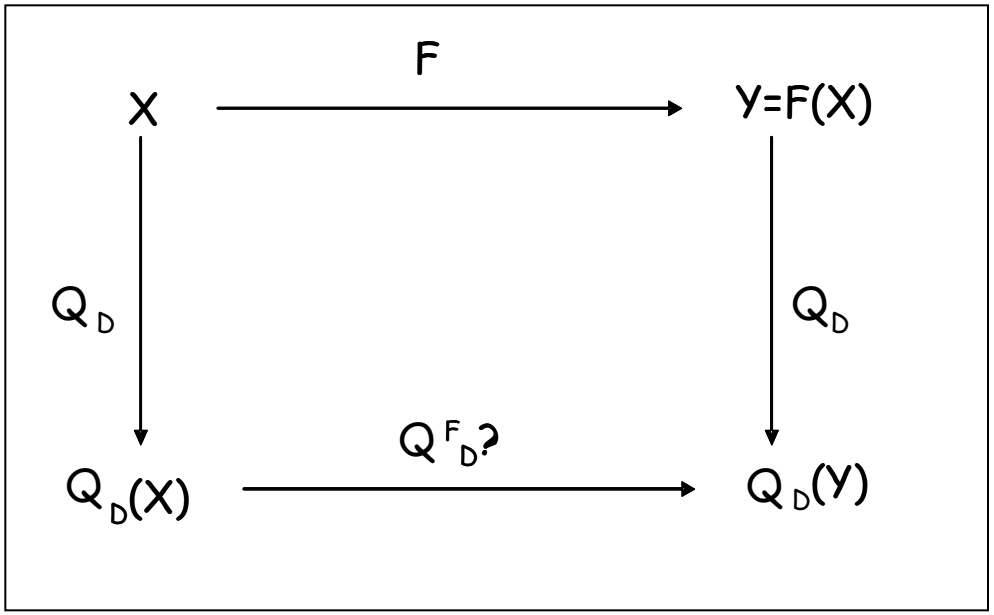
C. Batini & M. Scannapieco
Data and Information Quality Book Figures

Chapter 7: Activities for Information Quality

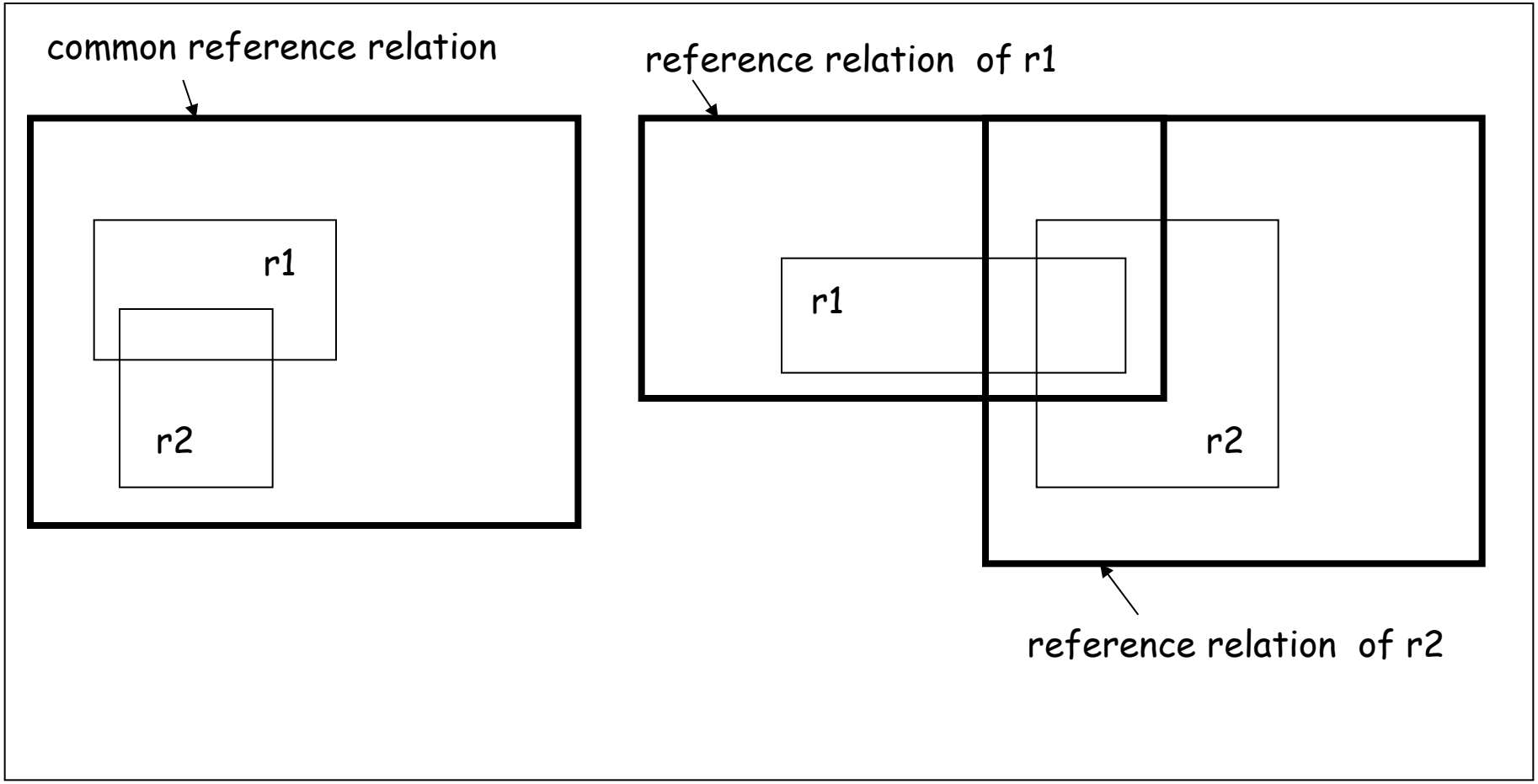
The general problem of quality composition



Comparison between approaches to quality composition

Paper	Model	Specific assumptions on the sources	Quality dimensions considered	Algebraic operators
Motro 1998	Relational model with OWA (implicit)	No assumption	Soundness Completeness	Cartesian Product Selection Projection
Parssiann 2002	Relational model with OWA (implicit)	Uniformely distributed errors in identifier attributes Error probabilites for all attributes independent of each other Uniformely distributed errors in non identifier attributes for mismember and other tuples	Accuracy Inaccuracy Mismembership Incompletness	Selection Projection Cartesian Product Join
Wang 2001	Relational model	Uniformly distributed errors	Accuracy	Selection Projection
Naumann 2004	Data integration system Set of data sources + Universal relation with CWA	Set relationships between sources - Disjointness - Quantified overlap - Independence (coincidential overlap) - Containment	Coverage Density Completeness	Join merge Full outer join merge Left outer join merge Right outer join merge
Scannapieco 2004	Relational model with OWA and CWA	Open world vs closed world assumption Set relationships between sources - Disjointness - Non quantified overlap - Containment	Completeness	Union Intersection Cartesian product

Assumptions for reference relations



Examples of accurate/inaccurate/mismember tuples and incomplete set in the Parssian approach

Id	LastName	Name	Role
1	Mumasia	John	Associate
2	Mezisi	Patrick	Full
3	Oado	George	Full
5	Ongy	Daniel	Full

(a) ideal relation *professor*

Id	LastName	Name	Role
1	Mumasia	John	Associate
2	Mezisi	Patrick	Full
3	Oado	Nomo	Full
4	Rosci	Amanda	Full

(b) real relation *professor*

Id	LastName	Name	Role
5	Ongy	Daniel	Full

(c) Set of incomplete tuples for *professor*

Symbols used in the exposition

Symbol	Meaning
r	input relation
$r_1, r_2, .., r_n$	a set of n input relations
s	output relation
| r |	size of the relation r
acc	accuracy
inacc	inaccuracy
cov	coverage
compl	completeness

Coverage composition functions in Naumann

Assumption/ operator	r_1 and r_2 disjoint	Quantified overlapping (= x)	r_1 contained in r_2
Join merge	0	|x| / |ur|	cov(r_1)
Left outer join merge	cov(r_1)	cov(r1)	cov(r_1)
Full outer join merge	cov(r_1) + cov(r_2)	cov(r_1) + cov(r_2) – |x| / |ur|	cov(r_1)

Examples of input relations

Id	LastName	Name	Role
1	Ongy	Daniel	Full
2	Mezisi	Patrick	Full
3	Oado	George	Full
4	Rosci	Amanda	Full

(a) dept1

Id	LastName	Name	Role
1	Mumasia	John	Associate
2	Mezisi	Patrick	Full
3	Oado	George	Full
4	Gidoy	Nomo	Associate
5	Rosci	Amanda	Full

(b) dept2

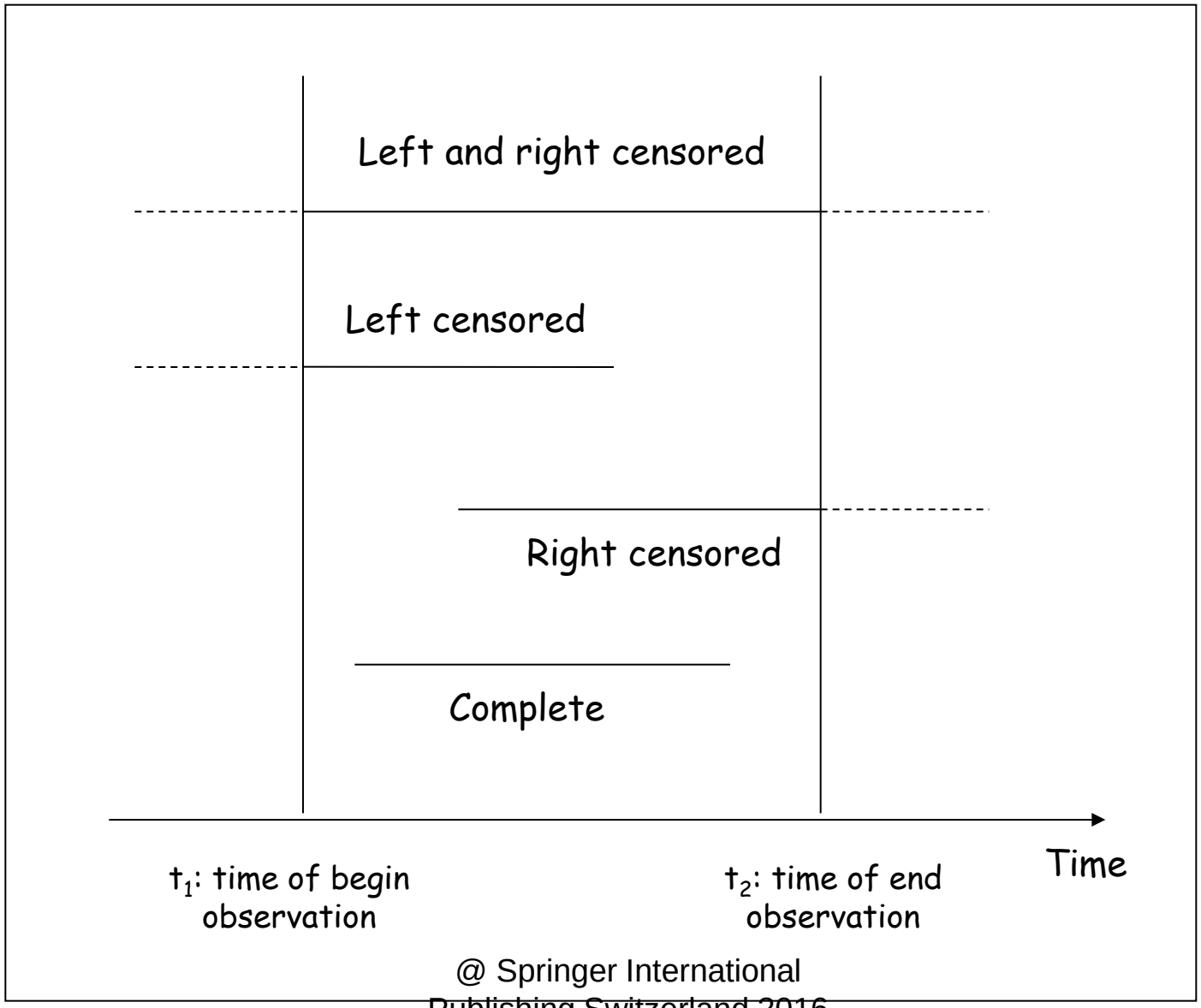
Id	LastName	Name	Role
1	Mumasia	John	Associate
2	Oymo	Vusi	Associate
3	Msgula	Luyo	Associate
4	Keyse	Frial	Associate

(c) dept3

Id	LastName	Name	Role
1	Ongy	Daniel	Full
2	Oado	George	Full

(d) dept4

Types of incomplete data in time series



Example of a control chart based on two attributes